Fundamentación teórica para la creación de un programa académico de ingeniería y ciencia de datos: una aplicación bibliométrica.
Theoretical foundation for the creation of an academic program of engineering and data science: a bibliometric application.

Frederick Andrés Mendoza-Lozano[1], Jose Wilmar Quintero-Peña[2], Oscar Leonardo Acevedo-Pabón[3], Jose Félix García-Rodriguez[4]
[1,2,3]*Institución Universitaria Politécnico Grancolombiano – Colombia*, [4]*Universidad Veracruzana – México*
ORCID: [1]*0000-0001-5087-4476*, [2]*0000-0002-6172-0453*, [3]*0000-0002-6172-0453*, [4]*0000-0002-7319-1472*

Recibido: 22 de junio de 2021.
Aprobado: 18 de agosto de 2021.

***Resumen*—** El objetico es definir un enfoque teórico entorno a la ciencia de datos, que incluya objeto de estudio y métodos, como paso previo para el diseño curricular de un programa académico. El texto inicia con una revisión de la literatura entorno a la evolución del concepto de dato y los fundamentos epistemológicos de la estadística y el análisis de datos, mediante el uso de algoritmos. Se continúa con la bibliometría de la producción científica de mayor relevancia, 2000 artículos, haciendo uso del enfoque de caracterización temática, mediante palabras clave tomadas de trabajos indexados en SCOPUS. Se encontró que la mayoría de las palabras clave y temáticas relevantes se refieren a los métodos de la modelación de datos con algoritmos y a la gestión de tecnología para la administración de grandes bases de datos. Se caracterizó la productividad del análisis de datos derivados de información textual, multimedia y la web. También se revelaron las temáticas referidas a las aplicaciones empresariales dirigidas a la gestión del conocimiento y la inteligencia de negocios. Se concluye que el concepto de dato, como objeto de estudio, se amplía gracias a los alcances del análisis de datos con algoritmos; este método se combina con el enfoque estadístico clásico, que provee modelos formales de mejor interpretación. Se concluyó que el campo de aplicación de la nueva ciencia de datos es bastante amplio, en particular cuando esta ciencia se utiliza en contextos interdisciplinarios. Lo anterior justifica el diseño curricular de un programa académico centrado en esta temática.

Palabras Claves: Ciencia de datos, Bibliometría, Minería de datos, Big data, Estadística clásica, Machine learning, Currículo**.**

***Abstract*—** The aim was to define a theoretical approach to data science, which includes object of study and methods, as a previous step for the curricular design of an academic program. The text begins with a review of the literature regarding the evolution of the concept of data and the epistemological foundations of statistics and data analysis, through the use of algorithms. The bibliometry of the most relevant scientific production continues, making use of the thematic characterization approach, using keywords taken from works indexed in SCOPUS. It was found that most of the relevant keywords and themes refer to the methods of data modeling with algorithms and the management of technology for the administration of large databases. The productivity of the analysis of data derived from textual, multimedia and web information was characterized. The themes related to business applications aimed at knowledge management and business intelligence were also revealed. The concept of data, as an object of study, is extended thanks to the scope of data analysis with algorithms; This method is combined with the classical statistical approach, which provides formal models of better interpretation. It was concluded that the field of application of the new data science is quite broad, particularly when this science is used in interdisciplinary contexts. The above justifies the curricular design of an academic program focused on this subject.

Keywords: Data science, Bibliometry, Data mining, Big data, Classical statistics, Machine learning, Curriculum.

*Autor para correspondencia.
Correo electrónico: famendoza@poligran.edu.co (Frederick Andrés Mendoza-Lozano).
La revisión por pares es responsabilidad de la Universidad de Santander.

Forma de citar: F. A. Mendoza-Lozano, J. W. Quintero-Peña, O. L. Acevedo-Pabón y J. F. García-Rodriguez, “Fundamentación teórica para la creación de un programa académico de ingeniería y ciencia de datos: una aplicación bibliométrica”, Aibi revista de investigación, administración e ingeniería, vol. 9, no. 3, pp. 49-58, 2021.

I. INTRODUCCIÓN

Durante el 2019, la Institución Universitaria Politécnico Grancolombiano (IUPG) decidió crear un nuevo programa de pregrado en el campo de la ciencia de datos. Una vez se obtenga el registro calificado por parte del Ministerio de Educación Nacional, será uno de los primeros pregrados en este campo de estudio, dado que un buen grupo de instituciones de educación superior ofertan programas similares a nivel de diplomado y posgrado.

Definir una nueva profesión en el país constituyó un reto, dada la necesidad de delimitar un objeto de estudio que combina contenidos curriculares de programas tradicionales como la estadística y la ingeniería de sistemas. De acuerdo con el decreto 1330 de 2019, que regula las condiciones de calidad de los programas de educación superior, en el desarrollo de los contenidos curriculares se deben exponer la fundamentación teórica en la que se soporta el objeto de estudio y los aspectos epistemológicos del campo de estudio. Lograrlo, se convirtió en el reto más notable de todo el trabajo de documentación, requerido para llevar a buen término la presentación de la solicitud del registro calificado.

Para estructurar la fundamentación teórica, se realizó una revisión de literatura sobre las perspectivas de investigación, de alcance interdisciplinario en el campo de la ciencia de datos. Se abordaron trabajos científicos publicados en revistas indexadas y se implementó el análisis bibliométrico. Con estos insumos se logró delimitar un campo de estudio y los métodos utilizados para analizar datos. Gracias a un enfoque histórico se ilustró la evolución desde la estadística clásica hasta la ciencia de datos. Este trabajo constituye la piedra angular de la argumentación que defiende la pertinencia de crear una nueva profesión, con denominación específica, cuyo perfil cuenta con un campo laboral amplio.

El presente artículo presenta los resultados de la investigación que se adelantó para fundamentar teóricamente el programa de ingeniería y ciencia de datos. Aquí se procede de la siguiente manera: iniciamos con una reflexión en torno a la evolución epistemológica de la ciencia de datos; luego proponemos algunas herramientas bibliométricas para describir la dinámica de la investigación en torno al análisis cuantitativo, tomando como insumo la producción científica indexada en SCOPUS. finalizamos presentando algunas conclusiones y sus implicaciones para la implementación curricular de este nuevo programa.

II. FUNDAMENTACIÓN TEÓRICA

a. La estadística a la ciencia de datos

En la década del sesenta, Tukey [1] propuso un enfoque distinto para la estadística como disciplina. En esencia, promueve centrar el objeto de estudio en el análisis de datos antes que en la modelación propia de la teoría estadística. Su postura era que las carencias de la teoría estadística y, en consecuencia, la demanda por un desarrollo interdisciplinario para el análisis de datos traería desarrollos extraordinarios en el corto plazo. Para su época esta propuesta fue polémica y disruptiva. Sin embargo, la evolución de las disciplinas con sólida base cuantitativa y, especialmente, los aportes de la ciencia de la computación al análisis de datos han mostrado que su visión fue acertada en términos generales: “Después de todo he llegado a sentir que mi interés central es el análisis de datos, en el cual incluyo entre otras cosas: procedimientos para analizar datos, técnicas para interpretar los resultados de esos procedimientos, caminos para planear la recolección de datos. Así, se hace más fácil su análisis y se otorga mayor precisión a toda la maquinaria y a los resultados (matemáticos) estadísticos que se aplican al análisis de datos” [1, pp 2].

La transición de la estadística a la ciencia de datos es un capítulo de la historia de la evolución de la ciencia que va desde de los modelos simples y estilizados para comprender la realidad hasta los marcos teóricos que dan cuenta de fenómenos de complejidad creciente [2]; por lo general, este tipo de fenómenos tiene que ver con procesos biológicos [3] o con sistemas sociales que evolucionan de forma no determinista, esto es, de acuerdo con decisiones individuales no predecibles [4]. En ese sentido, la creación de una ciencia de datos es el resultado de la combinación del enfoque clásico con el enfoque algorítmico. El primero asume que los datos se generan de acuerdo con un modelo estocástico; el segundo no se centra en teorizar sobre el mecanismo abstracto generador de datos, sino en extraer información relevante acerca de las relaciones entre las variables y las observaciones, así como en desarrollar instrumentos predictivos. Además, esta nueva ciencia encontró un objeto de estudio propio derivado de la evolución del concepto de dato.

Las raíces del método científico cartesiano crearon un mecanismo confiable para investigar la realidad a través de un procedimiento que garantiza resultados fiables, independientes del investigador. En ese sentido, Descartes [5] planteó la idea de separabilidad de las partes con el fin de estudiar los mecanismos de la realidad, a través del análisis de componentes simples y aislados. Más adelante, la epistemología positivista magnificó los resultados de la ciencia, al sostener que eran los únicos que tenían validez.

El conocimiento científico se caracteriza por ser racional, sistemático, exacto, verificable y, a su vez, falible, en tanto que no es dogmático. El desarrollo de la ciencia es una tensión constante entre la ciencia formal y la ciencia fáctica [6]. Buena parte del trabajo científico se ocupa de la estructuración de modelos que aspiran a representar la realidad, pero su validez o refutación dependen de la comprobación empírica que implica diseñar un experimento controlado en un laboratorio o hacer mediciones directamente en los fenómenos que no se pueden reproducir artificialmente.

Como consecuencia del trabajo científico experimental, emergió un cuerpo de conocimientos asociado a los errores de medición, que se presentan al menos en dos sentidos. De una parte, se debe lidiar con el asunto de la precisión, que a su vez está asociado a las tecnologías para el diseño experimental [7] y, de otra parte, está la pregunta por la validez científica basada en la evidencia empírica. Es decir, la extrapolación del hallazgo científico a toda una realidad, cuando este se fundamenta en los resultados de la observación de una parte aislada. Frente a esas dos objeciones, el trabajo científico desarrolló un conjunto de teorías asociadas a la recolección y análisis de datos. Así, se consolidó un grupo creciente de instrumentos estadísticos alineados con la brújula del pensamiento científico: modelos estilizados, simples y elegantes, creados bajo un procedimiento racional, con la rigurosidad del lenguaje matemático.

Los instrumentos estadísticos asociados a la medición e interpretación de datos numéricos apalancaron un avance extraordinario en la ciencia. Si bien el nacimiento de la estadística se remonta a una época reciente de la historia de la humanidad -hace unos dos siglos, como

respuesta a la necesidad de analizar los datos de los censos [8]-, casi todas las disciplinas modernas utilizan y desarrollan conceptos nucleares basados en la estadística.

Como se indicó antes, uno de los problemas más relevantes del trabajo científico es la posibilidad de incurrir en errores. Este asunto se deriva de la imposibilidad de tener mediciones absolutamente precisas y tiene estrecha relación con las simplificaciones propias de la investigación científica. Las comprobaciones empíricas siempre deben enfrentar el efecto de las imprecisiones de los instrumentos de medición, las variables omitidas en cualquier estudio cuantitativo, y un marco teórico analítico que simplifica tanto como le sea posible, a costa de pérdida de precisión [9]. Además, no se puede despreciar el hecho de que la mayoría de los fenómenos de los que se ocupa la ciencia son no deterministas. Aun cuando las teorías se pueden comprobar en laboratorio, los datos de las mediciones no son siempre los mismos, pues cambian en función de condiciones no controlables durante la experimentación y reflejan las variaciones en el tiempo; en buena medida, la ciencia y los instrumentos estadísticos se diseñaron para poder explicar esas variaciones.

Debido a su naturaleza indeterminada, muchos fenómenos de la realidad tuvieron que abordarse con el rigor que garantizaba que la ciencia podía avanzar con niveles altos de certeza, incluso siendo imposible deshacerse de los errores. Esa necesidad de certeza motivó el desarrollo de una completa teoría de probabilidades, con la cual la comprobación empírica consolidó un lenguaje propio. Es así como, con base en los modelos de la ciencia formal, se descartan escenarios que contradicen la teoría en favor de alternativas que la favorecen. En suma, el discurso científico propone que algo tiene validez porque es muy poco probable que sea de otra manera. Esto supone dos etapas y dos herramientas: en una etapa comprensiva, con los instrumentos de la teoría estadística se analizan las causas probables de las variaciones en las mediciones; en una etapa prospectiva, con los instrumentos que se diseñan con base en el cálculo de probabilidades se investigan los escenarios futuros en ejercicios de pronóstico.

De otro lado, el método científico encontró espacio en las ciencias sociales, donde se ha desarrollado ampliamente hasta nuestros días. Esa situación es problemática en la medida en que el traslado de los instrumentos y procedimientos de las ciencias básicas no siempre se ajusta de forma adecuada a la complejidad de los fenómenos sociales. No obstante, tomadas todas las previsiones del caso, se han consolidado campos específicos de aplicación que perduran hasta nuestra época. Es así como hoy se cuenta con una teoría cuantitativa del comportamiento humano, que se pone a prueba mediante la experimentación científica y su trabajo suele inscribirse en el área denominada psicometría. Otro ejemplo es el estudio de los fenómenos económicos a través de la econometría.

Al igual que la ciencia clásica o convencional, la teoría estadística ha sido criticada por sus limitaciones para dar cuenta de la realidad. El problema reside en la excesiva simplificación de los modelos analíticos en contraste con la diversidad y la complejidad de los fenómenos que se estudian. Esa simplicidad le otorga al investigador mayor capacidad para explicar fenómenos pasados, pero limita su capacidad predictiva [10]. En la teoría estadística, la prevención se manifiesta frente a la falta de ajuste de los datos reales a los modelos ideales en los que trabajan los estadísticos más puros. En esa forma de investigar prevalecen la especificidad del modelo, sus propiedades asintóticas y el desarrollo de la arquitectura matemática. Lo anterior no está libre de discusión pues, no siempre las conclusiones científicas explican relaciones causales ni se presentan interpretaciones falaces en la estadística utilizada para validar hipótesis [11]–[14].

La tendencia a centrarse en el modelo que se asume como el "generador de datos", antes que, en los datos reales es muy acentuada. Por ello, Breiman [9] la cataloga como una "cultura", entre los estadisticos puros, que se traslada a las aplicaciones interdisciplinarias en la econometría [15]–[17], la sociología [18]–[21], la psicología [22], [23] o las ciencias biológicas [24], [25]. Como alternativa, el mismo autor postula que existe otra "cultura" de naturaleza interdisciplinaria, cimentada en el análisis de datos con algoritmos. Este enfoque permite descubrir, con mayor precisión predictiva, el comportamiento de una variable de interés (dependiente) en función de un conjunto de variables explicativas (independientes), a través de un ejercicio computacional, complejo, preciso, y susceptible de ser optimizado mediante interacciones, aunque mucho más difícil de interpretar, con respecto a los instrumentos de la "cultura de los modelos generadores de datos".

Los nuevos enfoques del análisis de datos muestran que, en la práctica, la estadística es insuficiente para abordar todos los problemas que involucra el procesamiento de datos, especialmente cuando se trata de un gran conjunto de variables con capacidad explicativa [26]. Por ende, la formación básica transversal a todas las profesiones ya no se puede definir en términos de la alfabetización estadística sino de la alfabetización en uso de datos [27]. Aunque el desarrollo de la estadística parece estar enmarcado en la construcción de la teoría científica interdisciplinaria, la propuesta para crear un pregrado, sostiene que sí existe una ciencia de datos que delimita un objeto de estudio y una profesión en la que convergen los métodos clásicos para el análisis de datos con las herramientas computacionales recientes [28].

Para comprender eso, se necesita ceder a la necesaria integración de los conocimientos que hacen borrosas las fronteras disciplinarias, comprender las demandas sociales de expertos en el trabajo específico de datos con múltiples aplicaciones; y tomar en cuenta que el concepto de "dato", como objeto de estudio, ha variado en el tiempo gracias al desarrollo de la ciencia computacional. La influencia disciplinar de las técnicas modernas de análisis de datos se ven reflejadas, por ejemplo, en le ciencia económica, donde las herramientas de machine learning juegan un rol crucial en la solución de preguntas que surgen con las predicciones. Si bien los métodos de machine learning son importantes, no son suficientes porque existen retos econométricos como los problemas de causalidad, la identificación de contrafactuales y el comportamiento económico [29].

Es decir, el conocimiento de la teoría juega un papel fundamental que, al combinarlo con instrumentos de la modelación de datos con algoritmos, genera un mejor entendimiento de problemas económicos que pueden involucran el uso de grandes volúmenes de información. Con datos a gran escala se ha generado una oportunidad para realizar mediciones económicas, incluso en países en desarrollo. Por ejemplo, Blumenstockm et.al [30] miden la pobreza a nivel individual, mediante información tomada de teléfonos inteligentes. De otro lado, [31] calculan en tiempo real la dinámica del empleo y el consumo en China. Los anteriores son solo algunos de los retos que enfrenta la medición económica. Por ello se puede afirmar que los métodos de la ciencia de datos podrían complementar las respuestas a preguntas de orden social. Es decir, la aplicación interdisciplinaria de la estadística sugiere que esta no debe inscribirse en un campo científico único; sino que se trata de un conjunto de teorías e instrumentos de gran importancia, transversales a muchos campos de estudio.

Si bien la estadística clásica desarrolló un primer enfoque, en el que se asume que los datos se generan de acuerdo con un modelo estocástico con la intención de poder explicar datos numéricos, la posibilidad de representar otro tipo de información en términos de expresiones

matemáticas amplía significativamente el concepto de "dato". Hoy en día el análisis de datos abarca el procesamiento de imágenes, texto, audio e información geográfica, o una combinación de las anteriores, a través de la extracción masiva de información de la web [32]. Gracias al desarrollo de las computadoras, un poema puede ser a la vez una obra artística que se valora con la intermediación de las emociones y el espíritu humano, o un conjunto de datos que se representa en forma matricial para su análisis computacional. Un razonamiento semejante se puede hacer respecto de las imágenes o los audios. Además, con las tecnologías para almacenar grandes volúmenes de datos, el problema del análisis cambió; pasó de ser de una naturaleza estática, en el que se recolectan datos en un periodo específico para un estudio delimitado en el tiempo, a una dinámica, en la que se deben analizar datos que se recolectan en tiempo continuo [33].

La ampliación del concepto de dato y el interés de estudiar fenómenos complejos, que involucran a la vez un conjunto amplio de variables y una generosa cantidad de observaciones, han derivado en una ciencia, con un cuerpo de conocimientos amplio y robusto que trasciende la modelación estadística. Esta nueva ciencia cuenta con un objeto de estudio propio que se soporta en la transdisciplinariedad y su procedimiento está inscrito en una remarcable incertidumbre. Esta nueva ciencia es la ciencia de datos. La ciencia de datos se ocupa de entender y pronosticar los fenómenos que se estudian con herramientas computacionales; amplía el concepto de dato (relaciones entre una variable dependiente y un conjunto amplio de variables explicativas), y hace prevalecer el componente fáctico sobre el formal.

Al centrar su foco sobre la explicación del comportamiento de los datos, sus herramientas se centran en la validación de los métodos explicativos: validación cruzada, puntaje de los modelos propuestos, optimización de indicadores de puntaje de la precisión del modelo, validación de la mejor estrategia de modelación en contrastación de resultados luego de correrlos todos. Los pasos del procedimiento de esta nueva ciencia se resumen en tres [8]: 1) recolección de información, 2) definición de un conjunto de instrumentos que compiten por explicar los datos y, finalmente, 3) una evaluación de tales candidatos para elegir aquel que ofrezca mayor precisión en los pronósticos. Es así como el perfil del científico de datos se amplía, pues dentro de sus intereses se incorporan nuevos temas como la visualización de datos, el aprendizaje estadístico, la construcción de historias con datos y la preservación de los datos [34].

De acuerdo con [8], las áreas de trabajo en Ciencia de Datos se pueden clasificar en seis:

1) Recolección, preparación y exploración: usualmente los expertos consideran que este trabajo ocupa el 80% del tiempo empleado en todo el proceso de análisis. Como se discutió antes, el concepto de dato se amplió y, en consecuencia, las técnicas para la recolección y preparación son diversas y especializadas. Por ejemplo, web scrapping, pubmed scrapping, procesamiento de imágenes, y carga y limpieza de archivos de texto, entre otros.

2) Representación inicial y transformación: abarca el trabajo de transformar y preparar los datos de forma que revelen la mayor cantidad de información. Incluye habilidades tanto en bases de datos como en representación matemática de datos no numéricos.

3) Computación con datos: la ciencia de datos es intensiva en la programación estadística. Actualmente los dos lenguajes más usados son Python y R. Cualquier profesional formado en estos temas requerirá desarrollar un conocimiento avanzado en una de estas herramientas, si bien en su desempeño profesional se encontrará con ambas. Adicionalmente, un gran volumen de datos obliga a diseñar y correr tanto experimentos que ahorren capacidad de cómputo como herramientas de computación en la nube. En esta fase de los trabajos en ciencia de datos también se define el flujo de datos, antes de entrar en la fase de modelación.

4) Visualización y presentación: a través de la estadística descriptiva, los datos limpios y organizados hacen emerger mucha información importante. Una buena presentación de cifras relevantes puede ser suficiente para consolidar un discurso convincente o motivar la dirección que toma una decisión fundamental. Por su puesto, el uso de información estadística ante un auditorio no especializado puede ser objeto de manipulación, al incurrir en sesgos o imprecisiones sobre causalidad, magnificación exagerada, omisión de información importante, entre otras formas de argumentación falaz.

5) Modelación: la modelación en ciencia de datos se puede inscribir en lo que [9] denominó “las dos culturas”. O bien, se utilizan los “modelos generativos” que asumen que lo datos se generan de acuerdo con un proceso estocástico; o bien, los modelos predictivos que se fundamentan en algoritmos de aprendizaje. Algunos experimentos pueden usar una combinación de modelos de ambos tipos, para explorar cuáles predicen mejor. En ciertas circunstancias, se puede sacrificar algo de precisión en virtud de la capacidad de interpretación de los modelos generativos [35].

6) Investigación: la ciencia de datos está en desarrollo creciente. Es común encontrar en la web concursos internacionales para convocar expertos alrededor de los problemas de predicción -antes señalados- que aún exhiben niveles pobres de precisión. Así mismo, es de gran interés encontrar algoritmos de modelos supervisados y no supervisados, que economizan capacidad de cómputo. En el apartado sobre el análisis bibliométrico, se verá que, al explorar los resultados de una base de datos de indexación con la palabra “data”, el volumen de producción alrededor de la analítica es extraordinariamente alto. Por ello, tanto la comunidad científica como el sector productivo están muy interesado en encontrar nuevas herramientas para estos temas.

Adicionalmente, la gestión de grandes volúmenes de datos plantea retos para la reglamentación del uso de información pública y el diseño de estadísticas oficiales orientadas a la toma de decisiones. En consecuencia, los aspectos legales constituyen un objeto de estudio de interés para el programa de estudios en ciencia de datos [36]. Una forma alternativa de visualizar las áreas de desempeño es comprender en forma diferenciada las áreas de trabajo de la ingeniería de datos y las de la ciencia de datos. Las primeras, fuertemente asociadas al almacenamiento, recolección y limpieza; las segundas, relacionadas con la investigación sobre los datos (Figura 1).

Tabla 1: División entre Ingeniería de Datos y Ciencia de Datos.

Ingenieros de Datos	**Científicos de Datos**
Procesar datos en bruto	Probar hipótesis
Funcionamiento de los datos en la tras escena	Otorgar resultados a los usuarios de negocio
Construir infraestructura para consolidar y enriquecer numerosos conjuntos de datos	Aplicar algoritmos de aprendizaje de máquinas y otras aproximaciones analíticas
Manejar el procesamiento de datos a gran escala	Develar hallazgos en grandes volúmenes de datos
Monitorear y mantener sistemas	Interpretar los resultados del análisis
Preparar datos para el análisis	Desarrollar análisis articulado con herramientas visuales

Fuente: Elaboración propia.

III. METODOLOGÍA DEL ANÁLISIS BIBLIOMÉTRICO

Comúnmente, el análisis bibliométrico se utiliza para caracterizar la producción bibliográfica de una temática específica. Estos trabajos cuantifican la producción publicada e indexada en bases de datos científicas. Dependiendo del interés del investigador, el trabajo se puede centrar en el impacto de las publicaciones, la exploración de los autores clave, las revistas más relevantes o el desarrollo de los subtemas y su relevancia. En la presente propuesta curricular, la bibliometría se centra en la última aplicación. Analizar el despliegue del análisis de datos, en la investigación de mayor impacto internacional, dará luces para la estructuración del plan de estudios y, de manera especial, para la actualización de los contenidos programáticos de las asignaturas.

a. Selección de una base de datos, palabras clave y venta de observación

Los dos referentes de indexación internacionales más relevantes son ISI Web of Knowledge y SCOPUS. Al parecer, en Colombia hay más familiaridad con esta última base y, además, la IUPG tiene acceso institucional. Los resultados que se presentan a continuación se lograron haciendo uso del paquete Bibliometrix, que corre sobre el software de acceso libre R. La base conceptual de este instrumento se encuentra en el trabajo de Aria y Cuccurullo [37].

La selección de las palabras de búsqueda es una tarea esencial para el desarrollo de un análisis bibliométrico. Este trabajo implica un análisis especializado de tesauros y la validación de expertos. Para desarrollar la bibliometría del análisis de datos, se hicieron búsquedas combinando varios términos clave, que se referían a las más grandes áreas de teoría estadística paramétrica y no paramétrica, con las denominaciones más relevantes de los modelos de minería de datos supervisados y no supervisados. Un rasgo importante de esta metodología es que la interpretación de los resultados trasciende el marco conceptual por medio del cual se producen los resultados, pues depende en gran medida de la experticia del investigador.

Al final, los resultados más claros se obtuvieron usando el término más genérico posible. El término de búsqueda fue “data”. Si bien este término es demasiado genérico y parece una selección extremadamente simple, sus resultados tienen buena interpretación y se corresponden con el propósito de este trabajo: conocer el desarrollo temático del análisis de datos, en toda la producción científica internacional, para usarlo como insumo en la estructura curricular.

b. Análisis de la estructura conceptual

El propósito de este análisis es identificar los subtemas de la temática principal y agruparlos por similitudes, de acuerdo con el criterio de co-ocurrencia de palabras clave [37]. Para lograrlo, se parte de una matriz que cruza todas las palabras clave con los documentos, de manera que se vuelve relevante la aparición conjunta de palabras en los documentos.

Las palabras clave de las publicaciones se disponen en una matriz X (palabras clave vs. documentos), en donde X_{ij} toma el valor el valor de 1 (si la palabra clave i se incluye en el documento j) o el valor de 0 (en caso contrario). A través de un Análisis de Correspondencia Múltiple (ACM), se construye un plano reducido a dos dimensiones en el que las palabras se representan más cerca, en función de la similitud de sus distribuciones [38]–[40]. El ACM permite tanto un análisis exploratorio sin asumir restricciones sobre los datos, como una interpretación sencilla en la que se establecen clústeres de palabras clave, según su posición en el plano factorial de dos dimensiones [41].

c. Métricas de centralidad y densidad

El análisis de temas clave a través del criterio de co-occurencia se puede visualizar como una red. De esta manera, los palabras clave agrupadas en clústeres, por medio del algoritmo k-means, conforman grupos que adquieren densidad, cuando hay una alta co-occurencia de palabras clave dentro de un mismo clúster. De acuerdo con Cobo et al. [38], esta métrica se interpreta como el nivel de desarrollo dentro de una temática. Por su parte, la centralidad mide el grado de interrelación de la palabra clave de una temática con palabras clave de otras temáticas.

El índice de equivalencia [38] se define como

$$e_{ij} = c_{ij}^2 / c_i c_j \qquad (1)$$

Donde c_{ij} es el número de documentos en los cuales dos palabras clave i y j co-ocurren. Y c_i y c_j representan el número de documentos en los que cada uno aparece.

A partir de un índice de equivalencia, Callon et al. [42] interpretan las co-ocurrencias de palabras clave como una red. Por consiguiente, definen dos métricas clásicas: 1) la centralidad, que puede ser definida así:

$$c = 10 * \sum e_{kh} \qquad (2)$$

Donde k es una palabra clave que pertenece a un tema, y h es una palabra clave que pertenece a otros temas, y 2) la densidad, que puede ser definida así:

$$d = 100(\sum e_{ij}/w) \qquad (3)$$

Donde i y j son palabras clave que pertenecen a un mismo tema, y w es el número total de palabras clave dentro del tema.

En un plano cartesiano, los cuadrantes se pueden representar así (en sentido horario, empezando por el cuadrante superior de la izquierda):

- El primero representa los temas de mayor desarrollo (alta densidad), que a su vez están aislados, es decir, son muy especializados (baja centralidad).
- El segundo representa temas altamente desarrollados y transversales: estos son los “motores” de la investigación.

- El tercero representa temáticas de baja centralidad y densidad, es decir, o son muy nuevos o están decayendo en relevancia. Aún así, se debe tener en cuenta que este análisis se hizo con los trabajos más relevantes según el criterio de selección de SCOPUS.
- Finalmente, el cuarto cuadante presenta los temas básicos (de baja densidad) y transversales.

d. *Evolución de las temáticas en el tiempo*

La importancia de un enlace temático puede ser medida por los elementos que tienen en comun los temas enlazados. De acuerdo con Cobo et al. [38], el índice de inclusión se define así: sea T^t el conjunto de temas detectados en el subperiodo t, donde $U \in T^t$ representan cada tema detectado en el superiodo t. Sea $V \in T^{t+1}$ cada tema detectado en el periodo $t+1$. Se dice que hay una evolución temática desde U hacia V si hay palabras clave que se presentan en ambas y están asociadas a las redes temáticas. Por lo tanto, V puede ser considerado un tema evolucionado desde U. Las palabras clave $k \in U \cap V$ son consideradas un nexo temático o un nexo conceptual, y su nivel de importancia está dado por:

$$I = \frac{\#(U \cap V)}{Min(\#U, \#V)} \qquad (4)$$

IV. RESULTADOS ANÁLISIS E INTERPRETACIÓN

a. *Búsqueda con el término "data" en SCOPUS*

Por la amplitud del concepto, la búsqueda arroja resultados de trabajos publicados desde 1825 hasta 2020. En agosto de 2019, esta búsqueda arrojó 10.734.739 publicaciones. Para seleccionar los documentos más relevantes, se utilizó la opción correspondiente de SCOPUS**.** La exportación de resultados permitió consolidar una base de datos con las dos mil publicaciones más relevantes en toda la historia. Tras la aplicación de este filtro, en la Figura 1 se presentan resultados desde 1970, aunque el volumen de publicaciones relevantes crece significativamente a partir del año 2000.

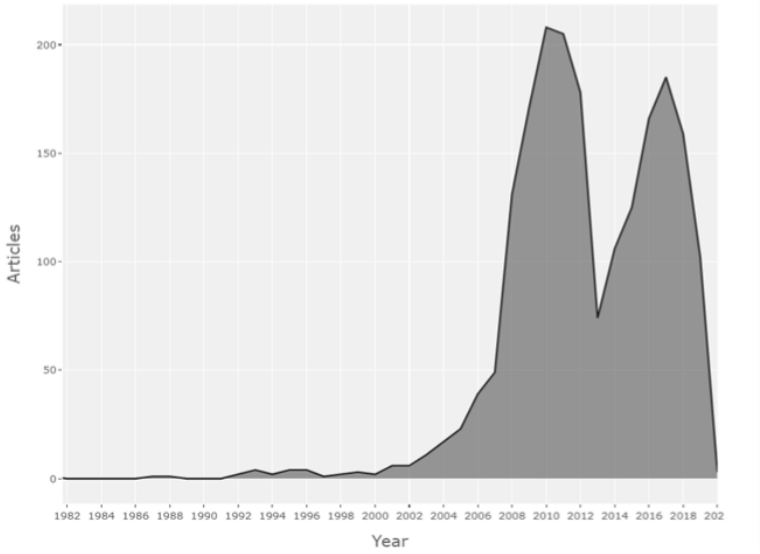

Figura 1: División entre Ingeniería de Datos y Ciencia de Datos.
Fuente: Elaboración propia.

En este conjunto de publicaciones relevantes, las palabras clave de autor de mayor importancia son "big data", "data mining". Ambos términos están a su vez en la lista de palabras clave de indexación, lo cual ya marca una tendencia para los temas de investigación más relevantes en este análisis de datos.

Tabla 2: Palabra clave de mayor frecuencia.

Palabras clave de autor	Artículos	Palabras clave de indexación	Artículos
Big data	208	Data mining	611
Data mining	169	Data handling	589
Data quality	124	Big data	465
Data integration	103	Information management	416
Data warehouse	103	Data reduction	326
Data management	87	Data warehouses	272
Data analysis	71	Data processing	252
Cloud computing	64	Data integration	250
Linked data	53	Data quality	235
Big data analytics	37	Data acquisition	232
Data cleaning	37	Data visualization	230
Data processing	37	Data sets	215
Data sharing	37	Digital storage	215
Data fusion	36	Metadata	205
Data model	36	Data structures	188

Open data	36	Data communication systems	168
EData warehousing	34	Visualization	167
Metadata	34	Data privacy	161
Data streams	33	Algorithms	148
Visualization	33	Database systems	143

Fuente: Elaboración propia.

SCOPUS utiliza un sofisticado procedimiento para evaluar la relevancia de las publicaciones. Una explicación sencilla acerca de cómo funciona se puede consultar en el portal de ayuda al usuario. En SCOPUS, hay dos tipos de palabras clave: en primer lugar, están las que seleccionan el o los autores; y, en segundo lugar, están las de indexación que seleccionan los proveedores de contenido Estas últimas son estandarizadas con base en un vocabulario público que restringe los términos que se pueden usar. A diferencia de las palabras clave de autor, las de indexación toman en cuenta sinónimos, diversas ortografías y plurales.

b. Estructura conceptual

El mapa revela información de cuatro clústeres de temáticas estrechamente relacionadas:

- Clúster 1: agrupa palabras clave asociadas a la investigación en minería de texto, análisis de texto, análisis geográfico y web scraping.
- Clúster 2: agrupa los temas nucleares en análisis de datos: almacenamiento, estructura de bases de datos, consulta, limpieza, analítica, visualización, computación en la nube, inteligencia artificial.
- Clúster 3: revela una cercanía, que también se puede identificar en el clúster 2, entre la minería de datos, los sistemas de comunicación, la seguridad de la información y los algoritmos de clasificación.
- Clúster 4: agrupa palabras clave que son de gran relevancia en ciencias administrativas. De esta manera, se refleja la relevancia de los temas de gestión de conocimiento, administración de sistemas de información, toma de decisión basada en datos.

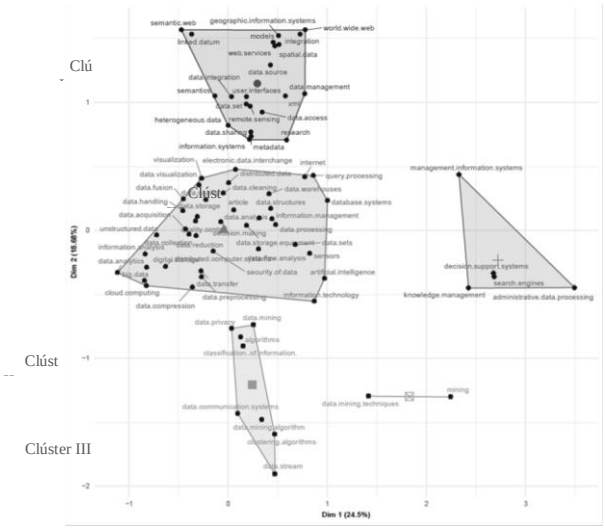

Figura 2: Clústeres de palabras clave.
Fuente: Elaboración propia.

c. Mapa temático

La Figura 3: presenta las temáticas según su nivel de centralidad y densidad. Los temas motores se refieren, en primer lugar, al procesamiento de bases de datos. Esto incluye: procesamiento, adquisición, calidad, control, reducción de dimensiones y, en una temática amplia, su análisis. Como temas en desuso, de baja centralidad y densidad, aparecen las temáticas asociados a la gestión de bodegas de datos y administración. En contraste, temas muy especializados y de alta densidad se refieren al análisis espacial, la semántica (modelamiento de bases de datos) y los metadatos. Por último, los temas transversales (más centrales que densos), que se podrían interpretar como interdisciplinarios, se refieren a la minería de datos, big data, técnicas de visualización y sistemas de comunicación.

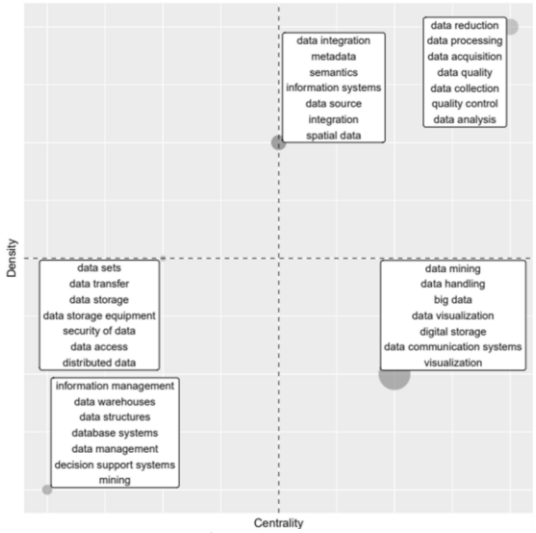

Figura 3: Mapa temático. Densidad vs. Centralidad.
Fuente: elaboración propia con datos de SCOPUS.

d. Evolución de las temáticas en el tiempo

La evolución temática refleja la relevancia de los instrumentos para la integración de datos en proyectos colaborativos de reúso. También se evidencian: 1) la consolidación del uso de herramientas de minería de datos en la inteligencia de negocios, 2) la migración del concepto de "administración de datos" hacia otro de mayor complejidad conocido como "gobierno de datos", y 3) en las migraciones temáticas es común que algunos temas de procesamiento básico sobre bases de datos migren hacia big data.

Tabla 3: Evolución temática.

Temática de origen 1970-2014	Temática de destino 2015-2020	índice de inclusión
data integration	data reuse	1
data mining	business intelligence	1
data management	open government data	0,5
data merging	big data	0,5
data mining	semi-structured data	0,25
data model	data Flow	0,25
data processing	big data	0,2
data processing	data fusión	0,2
data cleaning	big data	0,17
data cleaning	data quality	0,17
Reliability	data management	0,17
cloud computing	data lifecycle	0,14
data integration	big data mining	0,14
data management	visualization	0,12
data mining	data models	0,125
data mining	visualization	0,125
data warehousing	data models	0,125

Fuente: elaboración propia.

V. CONCLUSIONES

La ciencia de datos constituye un campo de investigación ampliado respecto de la estadística clásica, en al menos dos aspectos: en primer lugar, su campo ya no se circunscribe únicamente al dato numérico como objeto de estudio, dado que con los avances computacionales otros tipos de información como el sonido y las imágenes se pueden analizar, al ser dispuestos en forma matricial; en segundo lugar, la bibliometría como método de investigación en este campo refleja que, al hacer una búsqueda por la palabra clave “data”, presente en una gran cantidad de trabajos indexados en SCOPUS, las temáticas pertenecen principalmente a lo que [9] denomina "la cultura de la modelación con algoritmos". Todo esto aparece como resultado del análisis de los trabajos más relevantes y se evidencia tanto en la organización conceptual como en la evolución temática.

El presente trabajo demostró que un nuevo objeto de estudio, constituido alrededor de los datos, tiene aplicaciones interdisciplinarias de gran impacto. Sin duda, la investigación cuantitativa bajo el concepto ampliado de "dato" encuentra nuevos problemas y métodos en otras disciplinas.

El estudio de las temáticas relevantes y su evolución en el tiempo constituye el insumo esencial para entregar un diseño curricular que cumpla con lo que solicita la regulación colombiana contenida en el Decreto 1330 de 2019. La descripción de las temáticas y el análisis de los niveles de impacto de cada una son un primer referente riguroso para establecer los campos de enseñanza y el énfasis del nuevo programa. No obstante, este estudio deberá complementarse con el análisis de los planes de estudio de un conjunto amplio y diverso de propuestas locales y extranjeras.

VI. REFERENCIAS

[1] J. W. Tukey, "The future of data analysis," Ann. Math. Stat., vol. 33, no. 1, pp. 1–67, 1962.
[2] C. Maldonado and N. A. Gómez Cruz, El mundo de las ciencias de la complejidad. Un estado del arte. Bogotá, Colombia: Universidad del Rosario, 2010.
[3] C. Merow et al., "What do we gain from simplicity versus complexity in species distribution models?," Ecography (Cop.)., vol. 37, no. 12, pp. 1267–1281, 2014, doi: 10.1111/ecog.00845.
[4] K. V Katsikopoulos, "Bounded rationality: the two cultures," J. Econ. Methodol., vol. 21, no. 4, pp. 361–374, 2014, doi: 10.1080/1350178X.2014.965908.
[5] R. Descartes, Discurso del método. Ediciones Colihue SRL, 2004.
[6] M. Bunge, "La ciencia: su método y su filosofía," 1978.
[7] M. Frické, "Big data and its epistemology," J. Assoc. Inf. Sci. Technol., vol. 66, no. 4, pp. 651–661, 2015, doi: 10.1002/asi.23212.
[8] D. Donoho, "50 years of data science," J. Comput. Graph. Stat., vol. 26, no. 4, pp. 745–766, 2017, doi: 10.1080/10618600.2017.1384734.
[9] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," Stat. Sci., vol. 16, no. 3, pp. 199–231, 2001, doi: 10.1111/j.1740-9713.2005.00129.x.
[10] K. Mardia and W. Gilks, "Meeting the statistical needs of 21st-century science," Significance, vol. 2, no. 4, pp. 162–165, 2005, doi: 10.1111/j.1740-9713.2005.00129.x.
[11] W. M. Briggs, "Everything wrong with p-values under one roof," Studies in Computational Intelligence, vol. 809. Springer Verlag, 340 E. 64th Apt 9A, New York, United States, pp. 22–44, 2019, doi: 10.1007/978-3-030-04200-4_2.
[12] T. Derrick, "The criticism of inferential statistics," Educ. Res., vol. 19, no. 1, pp. 35–40, 1976.
[13] J. R. Jamison, "The use of inferential statistics in health and disease: a warning," South African Med. J., vol. 57, no. 19, pp. 783–785, 1980.
[14] B. L. Hopkins, B. L. Cole, and T. L. Mason, "A critique of the usefulness of inferential statistics in applied behavior analysis," Behav. Anal., vol. 21, no. 1, pp. 125–137, 1998.
[15] A. Charpentier, E. Flachaire, and A. Ly, "Econometrics and machine learning," Econ. Stat., vol. 2018, no. 505–506, pp. 147–169, 2018, doi: 10.24187/ecostat.2018.505d.1970.
[16] D. Qin, "Let's take the bias out of econometrics," J. Econ. Methodol., vol. 26, no. 2, pp. 81–98, 2019, doi: 10.1080/1350178X.2018.1547415.
[17] S. Athey and G. W. Imbens, "Machine Learning Methods That Economists Should Know about," Annu. Rev. Econom., vol. 11, pp. 685–725, 2019, doi: 10.1146/annurev-economics-080217-053433.
[18] M. Molina and F. Garip, "Machine Learning for Sociology," Annual Review of Sociology, vol. 45. Annual Reviews Inc., Department of Sociology, Cornell University, Ithaca, NY 14853, United States, pp. 27–45, 2019, doi: 10.1146/annurev-soc-073117-041106.
[19] S. Mützel, "Facing big data: Making sociology relevant," Big Data Soc., vol. 2, no. 2, p. 2053951715599179, 2015.
[20] D. A. McFarland, K. Lewis, and A. Goldberg, "Sociology in the era of big data: The ascent of forensic social science," Am. Sociol., vol. 47, no. 1, pp. 12–35, 2016.
[21] K. Healy and J. Moody, "Data visualization in sociology," Annu. Rev. Sociol., vol. 40, pp. 105–128, 2014.
[22] P. Barrett, "What if there were no psychometrics? Constructs, complexity, and measurement," J. Pers. Assess., vol. 85, no. 2, pp. 134–140, 2005, doi: 10.1207/s15327752jpa8502_05.
[23] N. Bolger, "Data analysis in social psychology," Handb. Soc. Psychol., vol. 1, pp. 233–265, 1998.
[24] D. Bzdok and J. P. A. Ioannidis, "Exploration, Inference, and Prediction in Neuroscience and Biomedicine," Trends Neurosci., vol. 42, no. 4, pp. 251–262, 2019, doi: 10.1016/j.tins.2019.02.001.
[25] A.-L. Boulesteix and M. Schmid, "Machine learning versus statistical modeling," Biometrical J., vol. 56, no. 4, pp. 588–593, 2014, doi: 10.1002/bimj.201300226.
[26] J. Wang and Q. Tao, "Machine learning: The state of the art," IEEE Intell. Syst., vol. 23, no. 6, pp. 49–55, 2008.
[27] R. Gould, "Data literacy is statistical literacy," Stat. Educ. Res. J., vol. 16, no. 1, pp. 22–25, 2017.
[28] P. Bühlmann, "Comments on: Data science, big data and statistics," Test, vol. 28, no. 2, pp. 330–333, 2019, doi: 10.1007/s11749-019-00646-6.
[29] S. Mullainathan and J. Spiess, "Machine learning: an applied econometric approach," J. Econ. Perspect., vol. 31, no. 2, pp. 87–106, 2017, doi: 10.1257/jep.31.2.87.
[30] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," Science (80-.)., vol. 350, no. 6264, pp. 1073–1076, 2015, doi: 10.1140/epjds/s13688-017-0125-5.
[31] L. Dong, S. Chen, Y. Cheng, Z. Wu, C. Li, and H. Wu, "Measuring economic activities of China with mobile big data," arXiv Prepr. arXiv1607.04451, 2016, doi: 10.1140/epjds/s13688-017-0125-5.
[32] B. Yu, "Embracing statistical challenges in the information technology age," Technometrics, vol. 49, no. 3, pp. 237–248, 2007, doi: 10.1198/004017007000000254.
[33] S. Tonidandel, E. B. King, and J. M. Cortina, "Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science," Organ. Res. Methods, vol. 21, no. 3, pp. 525–547, 2018, doi: 10.1177/1094428116677299.
[34] B. Beaton, A. Acker, L. Di Monte, S. Setlur, T. Sutherland, and S. E. Tracy, "Debating data science: A roundtable," Radic. Hist. Rev., vol. 2017, no. 127, pp. 133–148, 2017, doi: 10.1215/01636545-3690918.
[35] D. V Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," Electron., vol. 8, no. 8, 2019, doi: 10.3390/electronics8080832.
[36] P. J. H. Daas, M. J. Puts, B. Buelens, and P. A. M. van den Hurk, "Big data as a source for official statistics," J. Off. Stat., vol. 31, no. 2, pp. 249–262, 2015, doi: 10.1515/JOS-2015-0016.
[37] M. Aria and C. Cuccurullo, "bibliometrix: An R-tool for comprehensive science mapping analysis," J. Informetr., vol. 11, no. 4, pp. 959–975, 2017, doi: 10.1016/j.joi.2010.10.002.
[38] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field," J. Informetr., vol. 5, no. 1, pp. 146–166, 2011, doi: 10.1016/j.joi.2010.10.002.
[39] V. Batagelj and M. Cerinšek, "On bibliographic networks," Scientometrics, vol. 96, no. 3, pp. 845–864, 2013, doi: 10.1007/s11192-012-0940-1.

[40] K. Börner, C. Chen, and K. W. Boyack, “Visualizing knowledge domains,” Annu. Rev. Inf. Sci. Technol., vol. 37, no. 1, pp. 179–255, 2003, doi: 10.1002/aris.1440370106.
[41] C. Cuccurullo, M. Aria, and F. Sarto, “Foundations and trends in performance management. A twenty-five years bibliometric analysis in business and public administration domains,” Scientometrics, vol. 108, no. 2, pp. 595–611, 2016.
[42] M. Callon, J. P. Courtial, and F. Laville, “Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry,” Scientometrics, vol. 22, no. 1, pp. 155–205, 1991.